Towards a General Theory of Infinite-Width Limits of Neural Classifiers

Eugene A. Golikov [1]

Abstract

Obtaining theoretical guarantees for neural networks training appears to be a hard problem in a general case. Recent research has been focused on studying this problem in the limit of infinite width and two different theories have been developed: a mean-field (MF) and a constant kernel (NTK) limit theories. We propose a general framework that provides a link between these seemingly distinct theories. Our framework out of the box gives rise to a discrete-time MF limit which was not previously explored in the literature. We prove a convergence theorem for it, and show that it provides a more reasonable approximation for finite-width nets compared to the NTK limit if learning rates are not very small. Also, our framework suggests a limit model that coincides neither with the MF limit nor with the NTK one. We show that for networks with more than two hidden layers RMSProp training has a non-trivial discrete-time MF limit but GD training does not have one. Overall, our framework demonstrates that both MF and NTK limits have considerable limitations in approximating finite-sized neural nets, indicating the need for designing more accurate infinite-width approximations for them.

1. Introduction

Despite neural networks' great success in solving a variety of problems, theoretical guarantees for their training are scarce and far from being practical. It turns out that neural models of finite size are very complex objects to study since they usually induce a non-convex loss landscape. This makes it highly non-trivial to obtain any theoretical guarantees for the gradient descent training.

However theoretical analysis becomes tractable in the limit of infinite width. In particular, (Jacot et al., 2018) showed that if weights are parameterized in a certain way then the continuous-time gradient descent on neural network parameters converges to a solution of a kernel method. The corresponding kernel is called a neural tangent kernel (NTK).

Another line of work studies a mean-field (MF) limit of the training dynamics of neural nets with a single hidden layer (Mei et al., 2018; 2019; Rotskoff & Vanden-Eijnden, 2019; Sirignano & Spiliopoulos, 2020; Chizat & Bach, 2018; Yarotsky, 2018). In these works a neural net output is scaled differently compared to the work on NTK.

In our work we address several questions arising in this context:

1. Which of these two limits appears to be a more reasonable approximation for a finite-width network?
2. Do the two above-mentioned limits cover all possible limit models for neural networks?
3. Is it possible to construct a non-trivial mean-field limit for a multi-layer network?

The paper is organized as follows. In Section 2 we provide a brief review of the relevant studies. In Section 3 we consider hyperparameter scalings that lead to non-trivial infinite-width limits for neural nets with a single hidden layer. Our analysis clearly shows that MF and NTK limits are not the only possible ones. Also, our analysis suggests a discrete-time MF limit which appears to be a more reasonable approximation for a finite-sized neural network than the NTK limit if learning rates are not very small. We stress the difference between this discrete-time MF limit and a continuous-time one described in previous works and prove a convergence theorem for it. In Section 4 we show that when a neural net has at least three hidden layers a discrete-time MF limit becomes vanishing. Nevertheless, training a network with RMSProp instead of a plain gradient descent leads to a non-trivial discrete-time MF limit for any number of layers.

2. Related work

NTK limit. In their pioneering work Jacot et al. (2018) considered a multi-layer feed-forward network parameter-

[1] Neural Networks and Deep Learning lab., Moscow Institute of Physics and Technology, Moscow, Russia. Correspondence to: Eugene A. Golikov <golikov.ea@mipt.ru>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

ized as follows:

$$f(\mathbf{x}; W_{1:L}) = d_{L-1}^{-1/2} W_L \phi(d_{L-2}^{-1/2} W_{L-1} \ldots \phi(d_0^{-1/2} W_1 \mathbf{x})), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{d_0}$, d_i is a size of the i-th layer and $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$. The weights are initialized as $W_{l,ij}^{(0)} \sim \mathcal{N}(0,1)$.

Jacot et al. (2018) have shown that training this model with a continuous-time gradient descent is equivalent to performing a kernel gradient descent for some specific kernel; they called this kernel a neural tangent kernel (NTK). This kernel is generally stochastic and evolves with time, however, as they prove, it converges to a steady-state deterministic kernel as $d_{1:L-1} \to \infty$.

Lee et al. (2019) have shown that the training dynamics of the network (1) stays close to the training dynamics of its linearized version in the limit of infinite width; the linearization is performed with respect to weights. They also show that this statement holds for the discrete-time gradient descent as long as the learning rates are sufficiently small.

Arora et al. (2019) provide a way to effectively compute the NTK for convolutional neural networks. They found that a kernel method with the NTK still performs worse than the corresponding finite-width CNN. At the same time, as was noted by Lee et al. (2019), the training dynamics in the NTK limit is effectively linear. Bai & Lee (2019) artificially created a situation where a linearized dynamics was not able to track the training dynamics in the limit of infinite width. These two works show that the NTK limit is not perfect in the sense that it can be far from a realistic finite-size neural net.

Mean-field limit. There is a line of works (Mei et al., 2018; 2019; Rotskoff & Vanden-Eijnden, 2019; Sirignano & Spiliopoulos, 2020; Chizat & Bach, 2018; Yarotsky, 2018) that consider a two-layer neural net of width d in a mean-field limit:

$$f(\mathbf{x}; \mathbf{a}, W) = d^{-1} \mathbf{a}^T \phi(W^T \mathbf{x}) = d^{-1} \sum_{r=1}^{d} a_r \phi(\mathbf{w}_r^T \mathbf{x}), \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^{d_0}$; the weights are initialized independently on the width d and d goes to infinity. Note the difference in scaling the output function between (2) and (1) for $L = 2$. In the present case any weight configuration can be expressed as a point measure in $(a, \mathbf{w})$-space $\mathbb{R}^{d_0+1}$:

$$\mu[\mathbf{a}, W] = d^{-1} \sum_{r=1}^{d} \delta_{a_r} \otimes \delta_{\mathbf{w}_r}.$$

A neural network is then expressed as an integral over the measure:

$$f(\mathbf{x}; \mathbf{a}, W) = \int a\phi(\mathbf{w}^T \mathbf{x}) \, \mu[\mathbf{a}, W](da, d\mathbf{w}). \tag{3}$$

The above-mentioned works show that when learning rates are appropriately scaled width d, a gradient descent dynamics turns into a continuous-time dynamics for the measure μ in $(a, \mathbf{w})$-space driven by a certain PDE as d goes to infinity. This evolution in the weight space also drives the evolution of the model f (see (3)).

Note that those works that study a limit behavior of the discrete-time gradient descent (Sirignano & Spiliopoulos, 2020; Mei et al., 2018; 2019) require the number of training steps to grow with d since they prove convergence to a continuous-time dynamics. In contrast, in our work we find a similar mean-field-type limit that converges to a discrete-time limit dynamics.

There are several attempts to extend the mean-field analysis to multi-layer nets (Sirignano & Spiliopoulos, 2019; Nguyen, 2019; Fang et al., 2019). However this appears to be highly non-trivial to formulate a measure evolution PDE similar to a single-hidden-layer case (see the discussion of difficulties in Section 3.3 of Sirignano & Spiliopoulos (2019)). In particular, Sirignano & Spiliopoulos (2019) rigorously constructed an iterated mean-field limit for a two-hidden-layer case. In contrast, the construction of Nguyen (2019) applies to any number of layers while not being mathematically rigorous. Fang et al. (2019) claim to find a way to represent a deep network as a sequence of integrals over a system of probability measures. Given this, the loss becomes convex as a function of this system of measures. However they do not consider any training process.

It also has to be noted that Nguyen (2019) applied a weight initialization with a non-zero mean for their experiments with scaling multi-layer nets. As we show in Section 4, if the number of hidden layers is more than two and initialization has zero mean (which is common in deep learning), a mean-field limit becomes trivial.

3. Training a one hidden layer net with gradient descent

Here we consider a simple case of networks with a single hidden layer of width d trained with GD. Our goal is to deduce how one should scale its training hyperparameters (learning rates and initialization variances) in order to converge to a non-trivial limit model evolution as $d \to \infty$. We say that a limit model evolution is non-trivial if the model neither vanishes, nor diverges, and varies over the optimization process. We formalize this notion later in the text. We investigate certain classes of hyperparameter scalings that lead to non-trivial limit models. We find both existing MF and NTK scalings, as well as a different class of scalings that lead to a model that does not coincide with either MF or NTK limit. Finally, we discuss the ability of limit models to approximate finite-width nets.

Consider a one hidden layer net of width d:

$$f(\mathbf{x}; \mathbf{a}, W) = \mathbf{a}^T \phi(W^T \mathbf{x}) = \sum_{r=1}^d a_r \phi(\mathbf{w}_r^T \mathbf{x}),$$

where $\mathbf{x} \in \mathbb{R}^{d_0}$, $W = [\mathbf{w}_1, \ldots, \mathbf{w}_d] \in \mathbb{R}^{d_0 \times d}$, and $\mathbf{a} = [a_1, \ldots, a_d]^T \in \mathbb{R}^d$. The nonlinearity $\phi(z) = [z]_+ - \alpha[-z]_+$ for $\alpha > 0$ is considered to be the leaky ReLU and applied element-wise. We consider a loss function $\ell(y, z)$ that is continuosly-differentiable with respect to the second argument. We also assume $\partial \ell(y, z)/\partial z$ to be positive continuous and monotonic $\forall y$. The guiding example is the standard cross-entropy loss. The data distribution loss is defined as $\mathcal{L}(\mathbf{a}, W) = \mathbb{E}_{\mathbf{x}, y \in D_{train}} \ell(y, f(\mathbf{x}; \mathbf{a}, W))$, where D_{train} is a train dataset sampled from the data distribution $\mathcal{D}$.

Weights are initialized with isotropic gaussians with zero means: $\mathbf{w}_r^{(0)} \sim \mathcal{N}(0, \sigma_w^2 I)$, $a_r^{(0)} \sim \mathcal{N}(0, \sigma_a^2)$ $\forall r = 1 \ldots d$. The evolution of weights is driven by the gradient descent dynamics:

$$\Delta \theta_r^{(k)} = \theta_r^{(k+1)} - \theta_r^{(k)} = -\eta_\theta \frac{\partial \mathcal{L}(\mathbf{a}^{(k)}, W^{(k)})}{\partial \theta_r},$$

where θ is either a or $\mathbf{w}$.

Initialization variances, σ_a^2 and σ_w^2, generally depend on d: e.g. $\sigma_a^2 \propto d^{-1}$ for He initialization (He et al., 2015). This fact complicates the study of the limit $d \to \infty$. To work around this, we rescale our hyperparameters:

$$\hat{a}_r^{(k)} = \frac{a_r^{(k)}}{\sigma_a}, \quad \hat{\mathbf{w}}_r^{(k)} = \frac{\mathbf{w}_r^{(k)}}{\sigma_w}, \quad \hat{\eta}_a = \frac{\eta_a}{\sigma_a^2}, \quad \hat{\eta}_w = \frac{\eta_w}{\sigma_w^2}.$$

The GD dynamics preserves its form:

$$\Delta \hat{\theta}_r^{(k)} = -\hat{\eta}_\theta \frac{\partial \mathcal{L}(W^{(k)}, \mathbf{a}^{(k)})}{\partial \hat{\theta}_r}.$$

At the same time, scaled initial conditions do not depend on d anymore: $\hat{a}_r^{(0)} \sim \mathcal{N}(0, 1)$, $\hat{\mathbf{w}}_r^{(0)} \sim \mathcal{N}(0, I)$ $\forall r = 1 \ldots d$.

By expanding gradients we get the following:

$$\Delta \hat{a}_r^{(k)} = -\hat{\eta}_a \sigma_a \sigma_w \mathbb{E}_{\mathbf{x}, y} \nabla_f^{(k)} \ell \; \phi(\hat{\mathbf{w}}_r^{(k),T} \mathbf{x}), \tag{4}$$

$$\Delta \hat{\mathbf{w}}_r^{(k)} = -\hat{\eta}_w \sigma_a \sigma_w \mathbb{E}_{\mathbf{x}, y} \nabla_f^{(k)} \ell \; \hat{a}_r^{(k)} \phi'(\hat{\mathbf{w}}_r^{(k),T} \mathbf{x}) \mathbf{x}, \tag{5}$$

$$\hat{a}_r^{(0)} \sim \mathcal{N}(0, 1), \quad \hat{\mathbf{w}}_r^{(0)} \sim \mathcal{N}(0, I) \quad \text{for all } r = 1 \ldots d, \tag{6}$$

where we have denoted $f_d^{(k)}(\mathbf{x}) = \sigma_a \sum_{r=1}^d \hat{a}_r^{(k)} \phi(\sigma_w \hat{\mathbf{w}}_r^{(k),T} \mathbf{x})$ and $\nabla_f^{(k)} \ell = \left. \frac{\partial \ell(y,z)}{\partial z} \right|_{z = f_d^{(k)}(\mathbf{x})}$. We have also used the fact that $\phi(\sigma z) = \sigma \phi(z)$ for ϕ being the leaky ReLU. We shall omit $\mathbf{x}, y$ in the expectation from now on.

Denote $\sigma = \sigma_a \sigma_w$. Assume hyperparameters that drive the dynamics are scaled with d:

$$\sigma \propto d^{q_\sigma}, \quad \hat{\eta}_a \propto d^{\tilde{q}_a}, \quad \hat{\eta}_w \propto d^{\tilde{q}_w}.$$

We call a set of exponents $(q_\sigma, \tilde{q}_a, \tilde{q}_w)$ ”a scaling”. Every scaling define a limit model $f_\infty^{(k)}(\mathbf{x}) = \lim_{d \to \infty} f_d^{(k)}(\mathbf{x})$. We want this limit to be non-divergent, non-vanishing and not equal to the initialization $f_d^{(0)}$ for any $k \geq 1$. We call such scalings and corresponding limit models non-trivial.

3.1. Analyzing non-triviality

We start with introducing weight increments:

$$\delta \hat{a}_r^{(k)} = \hat{a}_r^{(k)} - \hat{a}_r^{(0)}, \quad \delta \hat{\mathbf{w}}_r^{(k)} = \hat{\mathbf{w}}_r^{(k)} - \hat{\mathbf{w}}_r^{(0)}.$$

Since our dynamics is symmetric with respect to permutation of indices r, we can assume the following:

$$|\delta \hat{a}_r^{(k)}| \propto d^{q_a^{(k)}}, \quad \|\delta \hat{\mathbf{w}}_r^{(k)}\| \propto d^{q_w^{(k)}}. \tag{7}$$

Our intuition here is that $|\delta \hat{a}_r^{(k)}| \sim d^{-1/2}$ for the NTK scaling, while $|\delta \hat{a}_r^{(k)}| \sim d^0$ for the MF scaling. We validate the assumption above numerically for some of the scalings in SM C. We proceed with decomposing the model:

$$\begin{aligned} f_d^{(k)}(\mathbf{x}) &= \sigma \sum_{r=1}^d (\hat{a}_r^{(0)} + \delta \hat{a}_r^{(k)}) \phi'(\ldots)(\hat{\mathbf{w}}_r^{(0)} + \delta \hat{\mathbf{w}}_r^{(k)})^T \mathbf{x} \\ &= f_{d,\emptyset}^{(k)}(\mathbf{x}) + f_{d,a}^{(k)}(\mathbf{x}) + f_{d,w}^{(k)}(\mathbf{x}) + f_{d,aw}^{(k)}(\mathbf{x}), \end{aligned} \tag{8}$$

where we define the decomposition terms as:

$$f_{d,\emptyset}^{(k)}(\mathbf{x}) = \sigma \sum_{r=1}^d \hat{a}_r^{(0)} \phi'(\ldots) \hat{\mathbf{w}}_r^{(0),T} \mathbf{x},$$

$$f_{d,a}^{(k)}(\mathbf{x}) = \sigma \sum_{r=1}^d \delta \hat{a}_r^{(k)} \phi'(\ldots) \hat{\mathbf{w}}_r^{(0),T} \mathbf{x},$$

$$f_{d,w}^{(k)}(\mathbf{x}) = \sigma \sum_{r=1}^d \hat{a}_r^{(0)} \phi'(\ldots) \delta \hat{\mathbf{w}}_r^{(k),T} \mathbf{x},$$

$$f_{d,aw}^{(k)}(\mathbf{x}) = \sigma \sum_{r=1}^d \delta \hat{a}_r^{(k)} \phi'(\ldots) \delta \hat{\mathbf{w}}_r^{(k),T} \mathbf{x}.$$

Here $\phi'(\ldots)$ is a shorthand for $\phi'((\hat{\mathbf{w}}_r^{(0)} + \delta \hat{\mathbf{w}}_r^{(k)})^T \mathbf{x})$.

Since all of the terms inside the sums are presumably power-laws of d (or at least do not grow or vanish with d), it is natural to assume power-laws for the decomposition terms also (see SM C for empiricial validation):

$$f_{d,\emptyset}^{(k)}(\mathbf{x}) \propto d^{q_{f,\emptyset}^{(k)}}, \; f_{d,a/w}^{(k)}(\mathbf{x}) \propto d^{q_{f,a/w}^{(k)}}, \; f_{d,aw}^{(k)}(\mathbf{x}) \propto d^{q_{f,aw}^{(k)}}$$

Here and later we will write ”a/w” meaning ”a or w”.

The introduced assumptions allow us to formulate the non-triviality condition in terms of the power-law exponents:

$$\max(q^{(k)}_{f,\emptyset}, q^{(k)}_{f,a}, q^{(k)}_{f,w}, q^{(k)}_{f,aw}) = 0 \;\forall k \geq 1; \tag{9}$$

$$\max(q^{(k)}_{f,a}, q^{(k)}_{f,w}, q^{(k)}_{f,aw}) = 0 \text{ or } q^{(k)}_{f,\emptyset} = 0 \text{ and } q^{(k)}_w \geq 0. \tag{10}$$

The first condition ensures that $\lim_{d\to\infty} f^{(k)}_d$ is finite and not uniformly zero, while the second one ensures that this limit does not coincides with the initialization (hence the learning dynamics does not get stuck as $d \to \infty$). In particular, the second condition requires either one of $f^{(k)}_{d,a}$, $f^{(k)}_{d,w}$, or $f^{(k)}_{d,aw}$ to contribute substantially to $f^{(k)}_d$ for large d, or, if the leading term is $f_{d,\emptyset}$, it requires $\lim_{d\to\infty} f^{(k)}_{d,\emptyset}$ not to coincide with $\lim_{d\to\infty} f^{(0)}_d$ (because $\phi'((\hat{\mathbf{w}}^{(0)} + \delta\hat{\mathbf{w}}^{(k)})^T\mathbf{x}) \nrightarrow \phi'(\hat{\mathbf{w}}^{(0),T}\mathbf{x})$ as $d \to \infty$ if $q^{(k)}_w \geq 0$).

In order to test Conditions (9) and (10), we have to relate the introduced q-exponents with the scaling $(q_\sigma, \tilde{q}_a, \tilde{q}_w)$. From the definition of decomposition (8) terms we get:

$$q^{(k)}_{f,\emptyset} = q_\sigma + \varkappa^{(k)}_\emptyset, \quad q^{(k)}_{f,a/w} = q^{(k)}_{a/w} + q_\sigma + \varkappa^{(k)}_{a/w},$$
$$q^{(k)}_{f,aw} = q^{(k)}_a + q^{(k)}_w + q_\sigma + \varkappa^{(k)}_{aw}, \tag{11}$$

where all $\varkappa \in \{1/2, 1\}$. We now use $q^{(k)}_{f,a}$ to illustrate where these equations come from. We have:

$$\mathbb{E}_{\,\hat{\mathbf{a}}^{(0)},\hat{W}^{(0)}} f^{(k)}_{d,a}(\mathbf{x}) = \sigma d \mathbb{E}\,\delta\hat{a}^{(k)}\phi'(\ldots)\hat{\mathbf{w}}^{(0),T}\mathbf{x} =$$
$$= \sigma d^{1+q^{(k)}_a} \mathbb{E}\,\frac{\delta\hat{a}^{(k)}}{d^{q^{(k)}_a}}\phi'(\ldots)\hat{\mathbf{w}}^{(0),T}\mathbf{x},$$

since all terms of the sum have the same expectation. Hence if the last expectation is non-zero in the limit of $d \to \infty$, then we have $\mathbb{E}\, f^{(k)}_{d,a}(\mathbf{x}) \propto \sigma d^{1+q^{(k)}_a}$ and consequently $q^{(k)}_{f,a}(\mathbf{x}) = q^{(k)}_a + q_\sigma + 1$; so, $\varkappa^{(k)}_a = 1$. However, if it is zero in the limit of $d \to \infty$, then we have to reason about the variance. We have $\mathbb{V}\mathrm{ar}\, f^{(k)}_{d,a}(\mathbf{x}) \propto \sigma^2 d^{1+2q^{(k)}_a}$ if all terms of the sum appear to be independent in the limit of $d \to \infty$, or $\mathbb{V}\mathrm{ar}\, f^{(k)}_{d,a}(\mathbf{x}) \propto \sigma^2 d^{2+2q^{(k)}_a}$ if they are perfectly correlated. Hence $q^{(k)}_{f,a}(\mathbf{x}) = q^{(k)}_a + q_\sigma + \varkappa^{(k)}_a$, where $\varkappa^{(k)}_a \in \{1/2, 1\}$. Generally, all $\varkappa$-terms can be defined if $q^{(k)}_a$ and $q^{(k)}_w$ are known.

We now relate $q^{(k)}_{a/w}$ with the scaling $(q_\sigma, \tilde{q}_a, \tilde{q}_w)$. First, we rewrite the dynamics in terms of weight increments:

$$\Delta\delta\hat{a}^{(k)}_r = -\hat{\eta}_a\sigma\mathbb{E}\,\nabla^{(k)}_f\ell\;\phi((\hat{\mathbf{w}}^{(0)}_r + \delta\hat{\mathbf{w}}^{(k)}_r)^T\mathbf{x}),$$
$$\Delta\delta\hat{\mathbf{w}}^{(k)}_r = -\hat{\eta}_w\sigma\mathbb{E}\,\nabla^{(k)}_f\ell\;(\hat{a}^{(0)}_r + \delta\hat{a}^{(k)}_r)\phi'(\ldots)\mathbf{x}, \tag{12}$$
$$\delta\hat{a}^{(0)}_r = 0,\; \delta\hat{\mathbf{w}}^{(0)}_r = 0,\; \hat{a}^{(0)}_r \sim \mathcal{N}(0,1),\; \hat{\mathbf{w}}^{(0)}_r \sim \mathcal{N}(0,I).$$

Recall that we are looking for scalings that lead to a non-divergent limit model $f^{(k)}_\infty$; hence $f^{(k)}_d$ should not grow with d. Then, since $\nabla^{(k)}_f\ell$ is strictly positive continuous and monotonic $\forall y$, we have $|\nabla^{(k)}_f\ell|$ bounded away from zero as a function of d. Also, since $\hat{a}^{(0)}_r \propto 1$ and $\|\hat{\mathbf{w}}^{(0)}_r\| \propto 1$, from the dynamics equations (12) we get:

$$q^{(1)}_a = \tilde{q}_a + q_\sigma, \quad q^{(1)}_w = \tilde{q}_w + q_\sigma, \tag{13}$$
$$q^{(k+1)}_a = \max(q^{(k)}_a, \tilde{q}_a + q_\sigma + \max(0, q^{(k)}_w)),$$
$$q^{(k+1)}_w = \max(q^{(k)}_w, \tilde{q}_w + q_\sigma + \max(0, q^{(k)}_a)).$$

The last two equations can be rewritten as:

$$q^{(k+1)}_{a/w} = \max(q^{(k)}_{a/w}, q^{(1)}_{a/w} + \max(0, q^{(k)}_{w/a})). \tag{14}$$

Here we have used the following heuristic rules:

$$u \propto d^{q_u},\; v \propto d^{q_v} \Rightarrow uv \propto d^{q_u+q_v},\; u+v \propto d^{\max(q_u,q_v)}.$$

Although these rules are not mathematically correct, we empirically validated the exponents predicted by equations (13) and (14): see SM C.

The $\varkappa$-terms together with equations (9), (10), (11), (13), and (14) define a set of sufficient conditions for a scaling $(q_\sigma, \tilde{q}_a, \tilde{q}_w)$ to define a non-trivial limit model. In the next section, we derive several solution classes for this system of equations. These classes contain both MF and NTK scalings, as well as a family of scalings that lead to a limit model that coincides with neither MF, nor NTK limits.

3.2. Non-trivial limits

Although deriving $\varkappa$-terms appears to be quite complicated generally, we derive them for several special cases.

Consider the case of $q^{(1)}_a < 0$ and $q^{(1)}_w < 0$. Equations (14) imply $q^{(k)}_a = q^{(1)}_a$ and $q^{(k)}_w = q^{(1)}_w$ $\forall k \geq 1$ then. We also conclude (see SM D) that $\varkappa^{(k)}_\emptyset = \varkappa^{(k)}_{aw} = 1/2$ and $\varkappa^{(k)}_{a/w} = 1$ in this case.

Conditions (9) and (10) then imply $q_\sigma \leq -1/2$ and $q^{(1)}_{a/w} \leq -1 - q_\sigma$ with an equality for at least one of $q^{(1)}_a$ or $q^{(1)}_w$. Because of the latter, and since in our case we have to have $\max(q^{(1)}_a, q^{(1)}_w) < 0$, we get a constraint $q_\sigma > -1$. Note also that in this case $q^{(k)}_{f,aw} < 0$.

Hence by taking $q_\sigma \in (-1, -1/2]$ and $\tilde{q}_{a/w} = q^{(1)}_{a/w} - q_\sigma \leq -1 - 2q_\sigma$ with at least one inequality being an equality, we define a non-trivial scaling. As a particular example of this case consider $q_\sigma = q^{(1)}_a = q^{(1)}_w = -1/2$. It follows than from (13) that $\tilde{q}_a = \tilde{q}_w = 0$. If we take $\hat{\eta}_a = \hat{\eta}_w = \hat{\eta}$ and $\sigma = \sigma^* d^{-1/2}$ then we get the following relations:

$$f_d^{(k)}(\mathbf{x}) = \sum_{r=1}^d \sigma^* d^{-1/2} \hat{a}_r^{(k)} \phi(\hat{\mathbf{w}}_r^{(k),T} \mathbf{x}),$$

$$\Delta\hat{\theta}_r^{(k)} = -\hat{\eta} \frac{\partial \mathcal{L}(W^{(k)}, \mathbf{a}^{(k)})}{\partial \hat{\theta}_r}, \quad \theta \in \{a, \mathbf{w}\},$$

$$\hat{a}_r^{(0)} \sim \mathcal{N}(0,1), \quad \hat{\mathbf{w}}_r^{(0)} \sim \mathcal{N}(0,I) \quad \text{for all } r = 1 \ldots d.$$

This system exactly corresponds to the one used in the NTK theory (Jacot et al., 2018) (see also eq. (1)).

Following (Jacot et al., 2018; Lee et al., 2019), we call a neural tangent kernel (NTK) the following function:

$$\Theta_d^{(k)}(\mathbf{x}, \mathbf{x}') = \sum_{r=1}^d \Bigg(\frac{\partial f^{(k)}(\mathbf{x})}{\partial \hat{a}_r} \frac{\partial f^{(k)}(\mathbf{x}')}{\partial \hat{a}_r} + \\ + \frac{\partial f^{(k)}(\mathbf{x})}{\partial \hat{\mathbf{w}}_r} \left(\frac{\partial f^{(k)}(\mathbf{x}')}{\partial \hat{\mathbf{w}}_r} \right)^T \Bigg) = \\ = \sigma^2 \sum_{r=1}^d \big(\phi(\hat{\mathbf{w}}_r^{(k),T} \mathbf{x}) \phi(\hat{\mathbf{w}}_r^{(k),T} \mathbf{x}') + \\ + \phi'(\hat{\mathbf{w}}_r^{(k),T} \mathbf{x}) \phi'(\hat{\mathbf{w}}_r^{(k),T} \mathbf{x}') \hat{a}_r^{(k),2} \mathbf{x}^T \mathbf{x}' \big).$$

If we consider training with the continuous-time GD this kernel drives the evolution of the model, see SM B:

$$\dot{f}_d^{(t)}(\mathbf{x}') = -\hat{\eta} \mathbb{E}_{\mathbf{x},y} \nabla_f^{(t)} \ell(\mathbf{x}, y)\, \Theta_d^{(t)}(\mathbf{x}, \mathbf{x}'), \qquad (15)$$

where we have taken $\hat{\eta}_a = \hat{\eta}_w = \hat{\eta}$.

For a finite d the NTK is a random variable, however when $\sigma \propto d^{-1/2}$, $\Theta_d^{(0)}$ converges to a deterministic non-degenerate limit kernel Θ_∞ due to the Law of Large Numbers. Moreover, if $\delta\hat{\mathbf{w}}_r^{(k)}$ and $\delta\hat{a}_r^{(k)}$ vanish with d (i.e. $q_{a/w}^{(k)} < 0$), then $\hat{\mathbf{w}}_r^{(k)}$ and $\hat{a}_r^{(k)}$ converge to $\hat{\mathbf{w}}_r^{(0)}$ and $\hat{a}_r^{(0)}$ respectively. Hence $\Theta_d^{(k)}$ converges to the same deterministic non-degenerate limit kernel Θ_∞.

However Θ_∞ becomes uniformly zero when $q_\sigma < -1/2$. Given $q_{a/w}^{(k)} < 0$, $\Theta_d^{(k)}$ still converges to $\Theta_\infty \equiv 0$. Nevertheless, if $\tilde{q} = \tilde{q}_a = \tilde{q}_w = -1 - 2q_\sigma$, then a new kernel $\tilde{\Theta}_d^{(k)} = \hat{\eta}\Theta_d^{(k)}$ converges to a non-vanishing deterministic limit kernel $\tilde{\Theta}_\infty$. The dynamics of the limit model is then driven by the above-mentioned limit kernel:

$$\dot{f}_\infty^{(t)}(\mathbf{x}') = -\mathbb{E}_{\mathbf{x},y} \nabla_f^{(t)} \ell(\mathbf{x}, y)\, \tilde{\Theta}_\infty(\mathbf{x}, \mathbf{x}'). \qquad (16)$$

Moreover, similar evolution equation holds also for the discrete-time dynamics, see again SM B:

$$f_\infty^{(k+1)}(\mathbf{x}') - f_\infty^{(k)}(\mathbf{x}') = -\mathbb{E}_{\mathbf{x},y} \nabla_f^{(k)} \ell(\mathbf{x}, y)\, \tilde{\Theta}_\infty(\mathbf{x}, \mathbf{x}'). \qquad (17)$$

Note also that if $q_\sigma < -1/2$ then the limit model vanishes at the initialization due to the Central Limit Theorem:

$$f_d^{(0)} = \sigma \sum_{r=1}^d \hat{a}_r \phi(\hat{\mathbf{w}}_r^T \mathbf{x}) \sim d^{q_\sigma + 1/2} \text{ for } d \to \infty.$$

We shall refer scalings for which $q_\sigma \in (-1, -1/2)$ as ”intermediate”. Since $f_\infty^{(0)}$ is zero for the intermediate scalings while it is not for the NTK scaling, limits induced by intermediate scalings do not coincide with the NTK limit. As we show in the next section, intermediate limits do not coincide with the MF limit either. Nevertheless, despite the altered initialization, the limit dynamics for intermediate scalings is still driven by the kernel similar to the NTK case: see eq. (17). Note that this ”intermediate” limit dynamics is the same for any $q_\sigma \in (-1, -1/2)$. Chizat et al. (2019) have already noted that taking $q_\sigma \in (-1, -1/2]$ leads to the so-called ”lazy-training” regime that in our terminology reads simply as $q_{a/w}^{(k)} < 0$.

3.2.1. Mean-field limit

If we take $q_a^{(1)} = q_w^{(1)} = 0$, then again, equations (14) imply $q_a^{(k)} = q_a^{(1)}$ and $q_w^{(k)} = q_w^{(1)}$ $\forall k \geq 1$. In this case we conclude that $\varkappa_\emptyset^{(k)} = \varkappa_{aw}^{(k)} = \varkappa_{a/w}^{(k)} = 1$ (see SM D).

Conditions (9) and (10) than imply $q_\sigma = -1$. It follows than from (13) that $\tilde{q}_a = \tilde{q}_w = 1$. Taking $\sigma = \sigma^* d^{-1}$ and $\hat{\eta}_{a/w} = \hat{\eta}^* d$ allows us to write the gradient descent step as a measure evolution equation.

Indeed, consider a weight-space measure: $\mu_d^{(k)} = \frac{1}{d} \sum_{r=1}^d \delta_{\hat{a}_r^{(k)}} \otimes \delta_{\hat{\mathbf{w}}_k^{(k)}}$. Given this, a neural network output can be represented as $f_d^{(k)}(\mathbf{x}) = \sigma^* \int \hat{a}\phi(\hat{\mathbf{w}}^T \mathbf{x})\, \mu_d^{(k)}(d\hat{a}, d\hat{\mathbf{w}})$ while the gradient descent step can be represented as $\mu_d^{(k+1)} = \mathcal{T}(\mu_d^{(k)};\ \eta^*, \sigma^*)$.

$\mu_d^{(0)}$ converges to $\mu_\infty^{(0)} = \mathcal{N}_{1+d_0}(0, I)$ in the limit of infinite width. Since η^* and σ^* are constants, the evolution of this limit measure is still driven by the same transition operator $\mathcal{T}$: $\mu_\infty^{(k+1)} = \mathcal{T}(\mu_\infty^{(k)};\ \eta^*, \sigma^*)$. In SM F we prove that than $\mu_d^{(k)}$ converges to $\mu_\infty^{(k)} = \mathcal{T}^k(\mu_d^{(0)})$ and $f_d^{(k)}$ converges to a finite $f_\infty^{(k)} = \sigma^* \int \hat{a}\phi(\hat{\mathbf{w}}^T \mathbf{x})\, \mu_\infty^{(k)}(d\hat{a}, d\hat{\mathbf{w}})$:

Theorem 1 (Informal version of Corollary 1 in SM F)**.** *If $\sigma \propto d^{-1}$, $\hat{\eta}_{a/w} \propto d$, and ℓ, ϕ, and the data distribution are sufficiently regular, then there exist limits in probability as $d \to \infty$ for $\mu_d^{(k)}$ and for $f_d^{(k)}(\mathbf{x})$ $\forall \mathbf{x}$ $\forall k \geq 0$.*

This theorem states the convergence of the discrete-time dynamics of a finite-width model to a discrete-time dynamics of a limit model. We call the corresponding limit model *a discrete-time mean-field limit.*

This limit differs from those considered in prior works. Indeed, previous studies on the mean-field theory resulted in a continuous-time dynamics for the limit model. For example, (Sirignano & Spiliopoulos, 2020) assume $\hat{\eta} \propto 1$. They prove that in this setup μ_d^{td} converges to a continuous-time measure-valued process ν^t for $t \in \mathbb{R}$. The limit process ν_t is driven by a certain integro-differentiable equation. In contrast, in our case $\mu_\infty^{(k)}$ is driven by a discrete-time process. Other works (e.g. Mei et al. (2018; 2019)) assume $\hat{\eta} = o(d)$ and also consider a continuous-time evolution for a limit model.

At the same time Rotskoff & Vanden-Eijnden (2019) and Chizat & Bach (2018) assume a learning rate scaling similar to ours but they consider a continuous-time gradient descent dynamics for the finite-width net.

Note that if $q_{a/w}^{(k)} < 0$ (as for NTK and intermediate scalings), then $\delta\hat{a}^{(k)}$ and $\delta\hat{\mathbf{w}}^{(k)}$ vanish as $d \to \infty$, hence $\mu_d^{(k)}$ converges to $\mu_\infty^{(0)} = \mathcal{N}_{1+d_0}(0, I)$. This means that in this case we cannot represent the dynamics of the limit model $f_\infty^{(k)}$ in terms of the dynamics of the limit measure, hence this case is out of the scope of the MF theory.

On the other hand, if $q_a^{(k)} = q_w^{(k)} = 0$, then a deterministic limit $\lim_{d\to\infty} \tilde{\Theta}_d^{(k)}(\mathbf{x}, \mathbf{x}')$ still exists due to the Law of Large Numbers, however this limit depends on step k since $\phi'(\hat{\mathbf{w}}^{(k),T}\mathbf{x}) \nrightarrow \phi'(\hat{\mathbf{w}}^{(0),T}\mathbf{x})$. Hence the dynamics of a limit model $f_\infty^{(k)}$ in the mean-field limit cannot be described in terms of a constant deterministic kernel.

So far we have considered two cases: $q_{a/w}^{(1)} < 0$ and $q_{a/w}^{(1)} = 0$. We elaborate other possible cases in SM E.

3.3. Infinite-width limits as approximations for finite-width nets

In the previous section we have introduced a family of scalings leading to different limit models. Limit models can be easier to study mathematically: for example, in the NTK limit the training process converges to a kernel method. If we show that a limit model approximates the original one well, we can substitute the latter with the former in our theoretical considerations.

Notice that conditions (9) and (10) allow some of (but not all of) $q_{f,\emptyset}^{(k)}$, $q_{f,a}^{(k)}$, $q_{f,w}^{(k)}$ and $q_{f,aw}^{(k)}$ to be less than zero. This means that corresponding terms of decomposition (8) vanish as $d \to \infty$. However for $d = d^*$, where $d^* < \infty$ is the width of a "reference" model, all of these terms are present. If we assume that indeed all of these terms obey power-laws with respect to d (which is a reasonable assumption for large d), then we can conclude that the fewer terms vanish as $d \to \infty$, the better the corresponding limit approximates the original finite-width net. We validate this assumption for the above-mentioned scalings in SM C.

One can see that for the NTK limit we have $q_{f,\emptyset}^{(k)} = q_{f,a}^{(k)} = q_{f,a}^{(k)} = 0$, hence the first three terms of decomposition (8) are preserved as $d \to \infty$, however $q_{f,aw}^{(k)} = -1$. In Figure 1 (center) we empirically check that this is indeed the case. One can notice however that the last term, which is not preserved, vanishes as $\hat{\eta} \to 0$ faster than $f_{d,a}^{(k)}$ and $f_{d,w}^{(k)}$. This reflects the fact that originally the NTK limit was derived for the continuous-time gradient descent for which the learning rate is effectively zero.

Note also that if $q_{a/w}^{(1)} < 0$ (for which the NTK scaling is a special case), then $q_{f,aw}^{(k)} < 0$ (see above), hence the last term of decomposition (8) always vanishes in this case. Hence the NTK scaling should provide the most reasonable approximation for finite-width nets among all scalings in this class. For comparison, we also consider the intermediate scaling $q_\sigma = -3/4$, $\tilde{q}_{a/w} = 1/2$ for which $q_{f,\emptyset}^{(k)} = -1/4$, $q_{f,a}^{(k)} = q_{f,w}^{(k)} = 0$, and $q_{f,aw}^{(k)} = -1/2$ for $k \geq 1$.

In contrast, for the MF limit we have $q_{f,\emptyset}^{(k)} = q_{f,a}^{(k)} = q_{f,w}^{(k)} = q_{f,aw}^{(k)} = 0$ for $k \geq 1$. Hence we expect all the terms of decomposition (8) to be preserved as $d \to \infty$. We check this claim empirically in Figure 1, center.

We also found it interesting to plot the case of the "default" scaling: $\sigma \propto d^{-1/2}$ and $\eta_{a/w} \propto 1$ (black curves). It corresponds to the situation when we make our network wider while keeping learning rates in the original parameterization constant. In this case $\hat{\eta}_a \propto d$, $\hat{\eta}_w \propto 1$, hence $\tilde{q}_a = 1$ and $\tilde{q}_w = 0$.

We compare final test losses for the above-mentioned scalings in Figure 1, left. As we see, all scalings except the default one result in finite limits for the loss while the default one diverges. As we see in Figure 1 (right), the mean-field limit tracks the learning dynamics of the reference network better than the other limits considered. It is interesting to note also that as the learning dynamics shows, MF and intermediate limits are deterministic while the NTK limit, as well as the reference model, are not. This is because the model at the initialization converges to zero for the first two cases. Also, this is the reason why the NTK limit becomes a better approximation for a finite-width net if learning rates are small enough (see Figure 3 in SM H). In this case the term $f_{d,aw}^{(k)}$, which is not preserved in the NTK limit, becomes negligible already for the reference network.

4. Training a multi-layer net

While reasoning about non-trivial limits for multi-layered nets is more technically involved, some qualitative results are still possible. For instance, we show that a (discrete-time) mean-field limit is vanishing for networks with more than three hidden layers. However, such a limit seems to

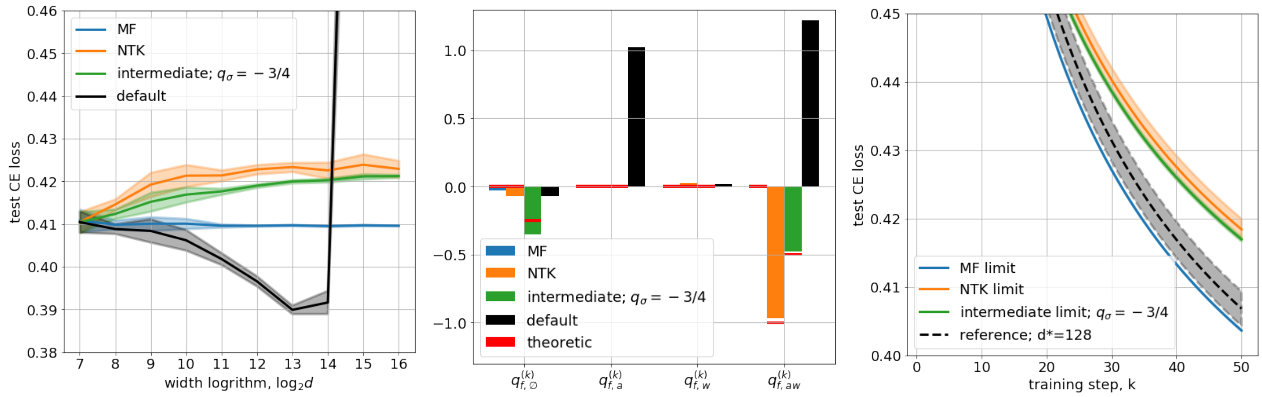

Figure 1. **MF, NTK and intermediate scalings result in non-trivial limit models for a single layer neural net. A limit model induced by the intermediate scaling differs from both MF and NTK limits.** *Left:* a final test cross entropy (CE) loss as a function of the width d. MF, NTK and intermediate scalings converge but the default scaling does not. The MF limit approximates the reference finite-width network better than all other limits. *Center:* numerical estimates for the exponents of the decomposition (8) terms as well as their theoretical values (denoted by red ticks). We see that for the default scaling some of the exponents are positive, hence corresponding decomposition terms diverge. For the MF limit all of the exponents are zeros, meaning all of the decomposition terms are preserved. Also, we see that our numerical experiments match the theory well. *Right:* the test CE loss as a function of training step k for the reference net and its limits. We see that 1) the MF limit best matches the reference, 2) the NTK limit is not deterministic while the intermediate limit is. This is because the model at the initialization converges to zero for the intermediate scaling. *Setup:* We train a 1-hidden layer net on a subset of CIFAR2 (a dataset of the first two classes of CIFAR10) of size 1000 with gradient descent. We take a reference net of width $d^* = 2^7 = 128$ trained with unscaled reference learning rates $\eta^*_a = \eta^*_w = 0.02$ and scale its hyperparameters according to MF (blue curves), NTK (orange curves), and intermediate scaling with $q_\sigma = -3/4$ (green curves, see text). We also make a plot for the case when we do not scale our learning rates (black curves) and scale standard deviations at the initialization as the initialization scheme of He et al. (2015) suggests. See SM A for further details.

exist if the network is trained with RMSProp.

Consider a multi-layered network with all hidden layers having width d:

$$f(\mathbf{x}; \mathbf{a}, V^{1:H}, W) = \sum_{r_H=1}^{d} a_{r_H} \phi(f^H_{r_H}(\mathbf{x}; V^{1:H}, W)), \tag{18}$$

where

$$f^{h+1}_{r_{h+1}}(\mathbf{x}; V^{1:h+1}, W) = \sum_{r_h=1}^{d} v^{h+1}_{r_{h+1} r_h} \phi(f^h_{r_h}(\mathbf{x}; V^{1:h}, W)),$$

$$f^0_{r_0}(\mathbf{x}, W) = \mathbf{w}^T_{r_0} \mathbf{x}.$$

Here again, all quantities are initialized with zero-mean gaussians: $a^{(0)}_{r_H} \sim \mathcal{N}(0, \sigma_a^2)$, $\mathbf{w}^{(0)}_{r_0} \sim \mathcal{N}(0, \sigma_w^2 I)$, and $v^{h,(0)}_{r_h r_{h-1}} \sim \mathcal{N}(0, \sigma_{v^h}^2)$.

We perform a gradient descent step for the parameters $\mathbf{a}$, $V^{1:H}$, W with learning rates η_a, $\eta_{v^{1:H}}$, and η_w respectively. We introduce scaled quantities in the similar manner as for the single hidden layer case:

$$\hat{a}^{(k)}_{r_H} = \frac{a^{(k)}_{r_H}}{\sigma_a}, \quad \hat{v}^{h,(k)}_{r_h r_{h-1}} = \frac{v^{h,(k)}_{r_h r_{h-1}}}{\sigma_{v^h}}, \quad \hat{\mathbf{w}}^{(k)}_{r_0} = \frac{\mathbf{w}^{(k)}_{r_0}}{\sigma_w},$$

$$\hat{\eta}_a = \frac{\eta_a}{\sigma_a^2}, \quad \hat{\eta}_{v^h} = \frac{\eta_{v^h}}{\sigma_{v^h}^2}, \quad \hat{\eta}_w = \frac{\eta_w}{\sigma_w^2}.$$

Given this, the gradient descent step on the scaled quantities writes as follows:

$$\Delta \hat{a}^{(k)}_{r_H} = -\hat{\eta}_a \sigma^{H+1} \mathbb{E}\, \nabla^{(k)}_f \ell(\mathbf{x}, y)\ \phi(\hat{f}^{H,(k)}_{r_H}(\mathbf{x})),$$

$$\Delta \hat{v}^{H,(k)}_{r_H r_{H-1}} = -\hat{\eta}_{v^H} \sigma^{H+1} \mathbb{E}\, \nabla^{(k)}_f \ell(\mathbf{x}, y)\ \hat{a}^{(k)}_{r_H} \phi(\hat{f}^{H-1,(k)}_{r_{H-1}}(\mathbf{x})),$$

$$\ldots$$

$$\Delta \hat{\mathbf{w}}^{(k)}_{r_0} = -\hat{\eta}_w \sigma^{H+1} \mathbb{E}_{\,\mathbf{x},y} \nabla^{(k)}_f \ell(\mathbf{x}, y) \times$$
$$\times \sum_{r_H=1}^{d} \hat{a}^{(k)}_{r_H} \phi'(\hat{f}^{H,(k)}_{r_H}(\mathbf{x})) \times$$
$$\times \sum_{r_{H-1}=1}^{d} \hat{v}^{H,(k)}_{r_H r_{H-1}} \phi'(\hat{f}^{H-1,(k)}_{r_{H-1}}(\mathbf{x})) \times \ldots$$
$$\ldots \times \sum_{r_1=1}^{d} \hat{v}^{2,(k)}_{r_2 r_1} \phi'(\hat{f}^{1,(k)}_{r_1}(\mathbf{x})) \hat{v}^{1,(k)}_{r_1 r_0} \phi'(\hat{\mathbf{w}}^{(k),T}_{r_0} \mathbf{x}) \mathbf{x}.$$

$$\hat{a}^{(0)}_{r_H} \sim \mathcal{N}(0,1), \ \hat{v}^{h,(0)}_{r_h r_{h-1}} \sim \mathcal{N}(0,1), \ \hat{\mathbf{w}}^{(0)}_{r_0} \sim \mathcal{N}(0, I), \tag{19}$$

where we have denoted $\hat{f}^{h,(k)}_{r_h}(\mathbf{x}) = f^h_{r_h}(\mathbf{x}; \hat{V}^{(k),1:h}, \hat{W}^{(k)})$ and $\sigma = (\sigma_a \sigma_{v^H} \ldots \sigma_{v^1} \sigma_w)^{1/(H+1)}$.

4.1. MF scaling leads to a trivial discrete-time MF limit

As we have noted in Section 3.2.1, the mean-field theory describes a state of a neural network with a measure in the weight space μ; similarly, it describes a networks' learning dynamics as an evolution of this measure. In particular, this means that weight updates cannot depend explicitly on the width d. Indeed, if they grow with d, then for some measure μ_∞ with infinite number of atoms this measure will diverge after a single gradient step. Similarly, if they vanish with d, then for a measure with an infinite number of atoms this measure will not evolve with gradient steps. Since we consider a polynomial dependence on d for our hyperparameters, our dynamics should not depend on d explicitly.

It is not obvious how to properly define a weight-space measure in the case of multiple hidden layers; the discussion in Section 3.3 of Sirignano & Spiliopoulos (2019); see also Fang et al. (2019). However, if we manage to define it properly, then each sum in the dynamics equation (19) will be substituted with an integral over the measure. Each such integral will contribute a d factor to the corresponding equation. Hence in order to have a learning dynamics independent on d we should have:

$$\hat{\eta}_{a/w}\sigma^{H+1}d^H \propto 1, \quad \hat{\eta}_{v^h}\sigma^{H+1}d^{H-1} \propto 1,$$

because there are H sums in the dynamics equation for $\hat{a}$ and $\hat{\mathbf{w}}$, and $H-1$ sums for $\hat{v}^{1:H}$. Similarly, since the network output should not depend on d, we should also have:

$$\sigma^{H+1}d^{H+1} \propto 1.$$

From this follows that $\sigma \propto d^{-1}$, $\hat{\eta}_{a/w} \propto d$, and $\hat{\eta}_{v^h} \propto d^2$.

As we show in SM G, for $H \geq 2$ this scaling leads to a vanishing limit: $f_d^{(k)}(\mathbf{x}) \to 0$ as $d \to \infty$ $\forall \mathbf{x}$ $\forall k \geq 0$. The intuition behind this result is simple: if $H \geq 2$ and $\phi(z) \sim z$ for $z \to 0$, then given the scaling above all of the weight increments vanish as $d \to \infty$ for $k = 0$. This means that the learning process cannot start in the limit of large d. We validate this claim empirically for $H = 2$ in Figure 2, center. In contrast, for the NTK scaling, which corresponds to $\sigma \propto d^{-1/2}$ and $\hat{\eta}_{a/v^h/w} \propto 1$, not all of the terms vanish. Nevertheless, if $H = 1$, a non-trivial mean-field limit seems to exist as our experiments demonstrate: see Figure 2, left.

Note that this result does not drive away the possibility of constructing a meaningful continuous-time MF limit, for which the limit dynamics is driven by an ODE. Also, we expect a meaningful MF limit to be possible for non-linearities that do not vanish near zero (e.g. for sigmoid).

4.2. Training a multi-layer net with RMSProp

Up to this point we have considered a GD training. Consider now training with RMSProp which updates the weights with normalized gradients. We show that in this case a mean-field limit exists and it is not trivial for any $H \geq 0$.

For the RMSProp training gradient updates look as follows:

$$\Delta\theta^{(k)} = \theta^{(k+1)} - \theta^{(k)} = -\eta_\theta \frac{\nabla_\theta^{(k)}}{\mathrm{RMS}_\theta^{(k)}}, \qquad (20)$$

where $\theta \in \{a_{r_H}, v^H_{r_H r_{H-1}}, \ldots, v^1_{r_1 r_0}, \mathbf{w}_{r_0}\}$. Here we have used following shorthands:

$$\nabla_\theta^{(k)} = \nabla_\theta \mathcal{L}(\Theta^{(k)}),$$

where $\Theta^{(k)} = \{a_{r_H}^{(k)}, v_{r_H r_{H-1}}^{H,(k)}, \ldots, v_{r_1 r_0}^{1,(k)}, \mathbf{w}_{r_0}^{(k)}\}$, and

$$\mathrm{RMS}_\theta^{(k)} = \sqrt{\sum_{k'=0}^{k} \beta^{k-k'} \nabla_\theta^{(k')} \odot \nabla_\theta^{(k')}} \quad \text{for } \beta \in (0,1).$$

Similarly to the GD case, we divide equation (20) by the standard deviation σ_θ of the initialization of the weight θ:

$$\Delta\hat{\theta}^{(k)} = -\frac{\eta_\theta}{\sigma_\theta} \frac{\nabla_{\hat{\theta}}^{(k)}}{\mathrm{RMS}_{\hat{\theta}}^{(k)}},$$

where $\nabla_{\hat{\theta}}^{(k)}$ and $\mathrm{RMS}_{\hat{\theta}}^{(k)}$ are defined similarly as above.

In this case we define scaled learning rates differently compared to the GD case: $\hat{\eta}_\theta = \eta_\theta / \sigma_\theta$.

As noted above, the mean-field analysis requires weight updates not to depend on d explicitly. Since our weight update rule uses normalized gradients, this condition reads simply as $\hat{\eta}_\theta \propto 1$ for all weights θ and $\sigma \propto d^{-1}$ since the model output $f[\mu_d; \mathbf{x}]$ should not depend on d explicitly.

Using similar reasoning as before (namely, weight increments should decay as $d^{-1/2}$) we can also define the NTK scaling: $\hat{\eta}_\theta \propto 1$ for all θ and $\sigma \propto d^{-1/2}$. We compare these two limits in Figure 2, right. Notice that similar to the single hidden layer case, the NTK limit preserves terms with low-order dependencies on learning rates (i.e. $f_{d,\emptyset}^{(k)}$, $f_{d,a/v^h/w}^{(k)}$), while the MF limit, being now non-vanishing, preserves terms with higher-order dependencies on them.

5. Conclusions

There are two different theories that study neural nets in the limit of infinite width: a mean-field theory and a kernel theory. These theories imply that if certain conditions are fulfilled, corresponding infinite-width limits are non-trivial, i.e. the resulting function neither explodes nor vanishes and the learning process does not get stuck as the width goes to infinity.

In our study we derive a set of sufficient conditions on the scaling of hyperparameters (weight initialization variances

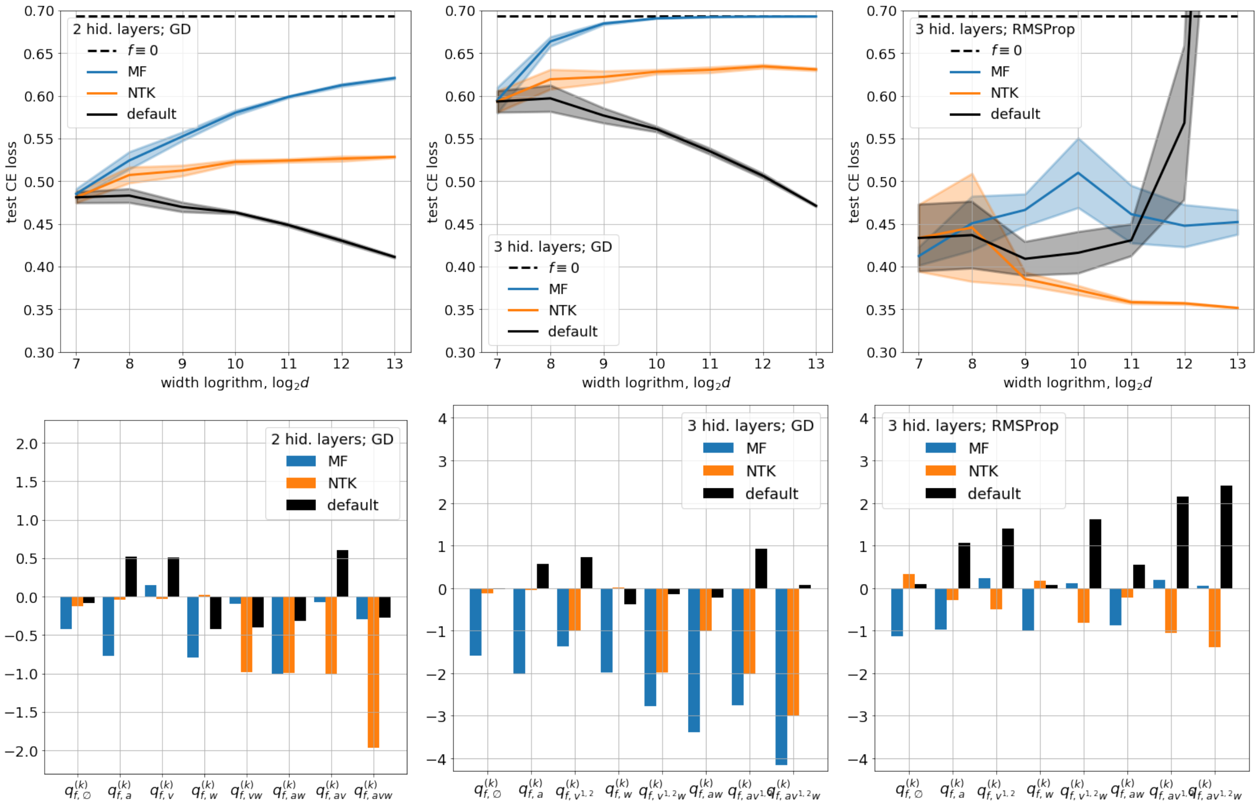

Figure 2. **MF and NTK limits for multilayer networks.** *Top row:* the final test cross-entropy (CE) loss as a function of width d. *Bottom row:* numerical estimates for exponents of terms of the decomposition of $f^{(k)}$, similar to eq.(8). All of these terms vanish for a network with (at least) three hidden layers in the MF limit, however, this is not the case when the number of hidden layers is two. Nevertheless, if we consider training with RMSProp, the MF limit becomes non-vanishing. For the NTK scaling, not all of the decomposition terms vanish in any case, however, some of them do, indicating possible discrepancies between the reference net and its NTK limit. *Setup:* We train a multi-layer net on a subset of CIFAR2 (a dataset of first two classes of CIFAR10) of size 1000 with either a plain gradient descent or RMSProp. We take a reference net of width $d^* = 2^7 = 128$ trained with (unscaled) reference learning rates $\eta_a^* = \eta_w^* = 0.02$ for GD and $\eta_a^* = \eta_w^* = 0.0002$ for RMSProp, and scale its hyperparameters according to MF (blue curves) and NTK (orange curves) scalings. We also make a plot for the case when we do not scale our learning rates (black curves) while scaling standard deviations at the initialization as the initialization scheme of He et al. (2015) suggests. See SM A for further details.

and learning rates) with width to ensure that we reach a non-trivial limit when the width goes to infinity. Solutions under these conditions include scalings that correspond to mean-field and NTK limits, as well as a family of scalings that lead to a limit model different from these two.

We propose a decomposition of our model and show that some of its terms may vanish for large width. We argue that a limit provides a more reasonable approximation for a finite-width net if as few of these terms vanish as possible.

Our analysis out of the box suggests a discrete-time MF limit which, to the best of our knowledge, has not been covered by the existing literature yet. We prove a convergence theorem for it and show that it provides a more reasonable approximation for finite-width nets than the NTK limit as long as learning rates are not too small.

As we show afterwards, a discrete-time mean-field limit appears to be trivial for a network with more than two hidden layers. Nevertheless, if we train our network with RMSProp instead of GD, the above-mentioned limit becomes non-trivial for any number of hidden layers.

Acknowledgments This work was supported by National Technology Initiative and PAO Sberbank project ID0000000007417F630002. We thank Mikhail Burtsev and Ivan Skorokhodov for valuable discussions and suggestions, as well as for help in improving the final version of the text.

References

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Pro-*

cessing Systems, pp. 8139–8148, 2019.

Bai, Y. and Lee, J. D. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pp. 3036–3046, 2018.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pp. 2933–2943, 2019.

Fang, C., Gu, Y., Zhang, W., and Zhang, T. Convex formulation of overparameterized deep neural networks. *arXiv preprint arXiv:1911.07626*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, pp. 8570–8581, 2019.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Mei, S., Misiakiewicz, T., and Montanari, A. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. *arXiv preprint arXiv:1902.06015*, 2019.

Nguyen, P.-M. Mean field limit of the learning dynamics of multilayer neural networks. *arXiv preprint arXiv:1902.02880*, 2019.

Rotskoff, G. M. and Vanden-Eijnden, E. Trainability and accuracy of neural networks: an interacting particle system approach. *stat*, 1050:30, 2019.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of deep neural networks. *arXiv preprint arXiv:1903.04440*, 2019.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

Yarotsky, D. Collective evolution of weights in wide neural networks. *arXiv preprint arXiv:1810.03974*, 2018.